Genomics 96 (2010) 290–302

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Tracing phylogenomic events leading to diversity of *Haemophilus influenzae* and the emergence of Brazilian Purpuric Fever (BPF)-associated clones

Leka Papazisi [a], Shashikala Ratnayake [a], Brian G. Remortel [a], Geoffrey R. Bock [a], Wei Liang [a], Alexander I. Saeed [a], Jia Liu [a], Robert D. Fleischmann [a], Mogens Kilian [b], Scott N. Peterson [a,*]

[a] *Pathogen Functional Genomics Resource Center (PFGRC), The J. Craig Venter Institute (JCVI), 9712 Medical Center Drive, Rockville, MD 20850, USA*
[b] *Institute of Medical Microbiology and Immunology, University of Aarhus, DK-8000 Aarhus C, Denmark*

ARTICLE INFO

Article history:
Received 30 January 2010
Accepted 14 July 2010
Available online 30 July 2010

Keywords:
Haemophilus
Brazilian Purpuric Fever
Pathogen emergence
Virulence
Comparative genomics
Microarray

ABSTRACT

Here we report the use of a multi-genome DNA microarray to elucidate the genomic events associated with the emergence of the clonal variants of *Haemophilus influenzae* biogroup aegyptius causing Brazilian Purpuric Fever (BPF), an important pediatric disease with a high mortality rate. We performed directed genome sequencing of strain HK1212 unique loci to construct a species DNA microarray. Comparative genome hybridization using this microarray enabled us to determine and compare gene complements, and infer reliable phylogenomic relationships among members of the species. The higher genomic variability observed in the genomes of BPF-related strains (clones) and their close relatives may be characterized by significant gene flux related to a subset of functional role categories. We found that the acquisition of a large number of virulence determinants featuring numerous cell membrane proteins coupled to the loss of genes involved in transport, central biosynthetic pathways and in particular, energy production pathways to be characteristics of the BPF genomic variants.

1. Introduction

Haemophilus influenzae (Hi) is a species of gram-negative, rod-shaped bacteria that have primarily evolved to live symbiotically in the upper respiratory tract of humans. The loss of genes for the synthesis of key metabolic compounds (heme and nicotineamide adenine dinucleotide, NAD) renders *H. influenzae* unable to survive in environments outside its human host. This species is comprised of distinct evolutionary lineages that express one of six different capsular serotypes, a through f, along with a highly diverse group of non-encapsulated bacteria displaying a non-clonal relationship [1,2].

Despite the commensal life-style of most *H. influenzae* strains, they remain one of the five most frequent causes of death among pathogenic microorganisms [3]. Prior to the introduction of wide-spread vaccination in many countries, *H. influenzae* strains expressing the serotype b capsule ranked among the three leading causes of bacterial meningitis worldwide. This remains true in many developing countries. It is estimated that at least 3 million cases of serious disease and an estimated 400,000–700,000 deaths occur in young children worldwide per year due to *H. influenzae* infection [3]. Invasive infections are primarily associated with strains representing one (Division I) of two distinct serotype b lineages, that display a characteristic partial duplication of the capsular biosynthesis (*cap*) locus [4]. Division II *H. influenzae* serotype b strains and those expressing any of the other five capsular serotypes are infrequent causes of infection. The non-encapsulated ("non-typeable") *H. influenzae* are regular inhabitants of the human pharynx, particularly during childhood, and are frequent causes of mucosal infections at sites contiguous with the upper respiratory tract, e.g. otitis media, sinusitis, lower respiratory tract infections, and conjunctivitis [1].

Occasionally, non-encapsulated *H. influenzae* strains cause invasive infections in the absence of apparent predisposing conditions in the patient. A striking example is the disease described as Brazilian Purpuric Fever (BPF), a fulminant life-threatening pediatric infection caused by particular clones of *H. influenzae* belonging to the biogroup aegyptius. BPF was first recognized in 1984, when 10 children in a Brazilian town of 20,000 persons died of an acute febrile illness associated with hemorrhagic skin lesions, septicemia, hypotensive shock, vascular collapse, and death. Alarmingly, the time of death from the time of symptom onset was within 48 h. The disease is characteristically preceded by purulent conjunctivitis that is resolved prior to the onset of fever. Additional outbreaks occurred subsequently in Brazil and cases with clinical symptomology indistinguishable from those of BPF have been described in Australia and the United States. These infections were caused by strains distinct from the Brazilian isolates and therefore considered to be the result of independent evolutionary events [5–10]. Common to all BPF isolates from such infections is the biotype III classification to, one of eight

* Corresponding author.
E-mail address: scottp@jcvi.org (S.N. Peterson).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.07.005

described *H. influenzae* biotypes [1]. In contrast to typical strains of *H. influenzae*, BPF isolates lack the ability to ferment xylose. It remains largely unknown how these bacteria evade the innate immune system and why they cause clinical attributes more commonly associated with meningococcal disease. Phylogenetic analysis indicated that the Brazilian BPF clones were most closely related to the group formally known as *H. aegyptius* [6,7], a recognized cause of an acute purulent and contagious form of conjunctivitis occurring as seasonal endemics, particularly in hot climates, including southern portions of the US. According to traditional taxonomic criteria, bacteria referred to as *H. influenzae*, *H. aegyptius*, and *H. influenzae* biogroup aegyptius are all members of a single species; however, there are formal obstacles to their unification into a single species. *H. aegyptius* isolates do not cause invasive disease, however they do share extensive genetic traits in common with the BPF clones.

The presence of a 24 MDa cryptic plasmid and the *H. influenzae* insertion sequence IS*1016* have been previously considered unique features of the BPF isolates [11,12]. However, this plasmid does not encode recognizable virulence factors, drawing the significance of this observation into question [13]. *H. influenzae* strain KW20/Rd, a capsule-deficient derivative of an essentially non-pathogenic serotype d strain, was the first living organism to have its complete genome sequence determined [14]. More recently, three complete and 11 additional draft genome sequences of non-encapsulated *H. influenzae* strains were reported [15–17]. Comparative genomics of *H. influenzae* genomes reveal various lineage-specific genes, including virulence factors. However, the underlying evolutionary processes leading to emergence and the genetic basis of invasive potential remain unclear [8,9,12,14–19].

In this report we demonstrate the utility of a multi-genome microarray for evaluating the gene content and extent of genomic diversity of 21 *H. influenzae* group members, including 14 *H. influenzae* strains, four *H. aegyptius*, two *H. influenzae* biogroup aegyptius and one *H. haemolyticus*. These strains represent a wide phenotypic diversity and geographic coverage of the species. Among the query group analyzed were three isolates obtained from patients associated with Brazilian Purpuric Fever (BPF), a fulminant fatal pediatric disease. The multi-genome-array, unlike a single-genome DNA microarray, reveals both gene loss and gene complement patterns with respect to any query genome and enabled us to identify evolutionary events associated with the emergence of BPF clones from *H. influenzae* variants associated with conjunctivitis.

2. Materials and methods

2.1. Bacterial strains and DNA techniques

The *Haemophilus* strains used in this study (Table 1) were grown and propagated as previously described [1]. All strains have been characterized extensively as part of a population genetic analysis [7] and were further analyzed by a multi-locus sequence typing (MLST) scheme developed for this species [2]. The majority of these strains were selected to represent major lineages with this taxon on the basis of a previously reported population genetic study that used both Multi

Table 1
The strains used in the study.

Strain name	Designation	Serotype	Country of origin	Biotype	Division	Virulence	Aberrant traits	Alterative designations	References
HK1368	*H. influenzae*	[b]	Denmark	I	I	Meningitis			
HK715	*H. influenzae*	[b]	USA	I	II (ET26)	Less virulent		1066/1971	[7]
HK635	*H. influenzae*	[c]	Papua N. Guinea	III		Lung aspirate		43/LA/Tarr	
HK2122	*H. influenzae*		Germany	III		Septicemia	Xylose negative		[69]
KW20/Rd	*H. influenzae*	Rough derivative of serotype d		IV				ATCC51907	[14]
HK2067	*H. influenzae*	[f]	USA			Septicemia		GA1463	[70]
HK1136	*H. influenzae*		USA	II				ATCC9134	
HK61	*H. influenzae*	NT[a]	Denmark	I		Pharynx	Phylogenetically distant (MLEE, MLST, *iga*)		[71]
HK1220	*H. influenzae*	NT	Brazil	III		Oropharynx, healthy subject		60/86	[7]
HK1210	*H. influenzae*	NT	Denmark	II		Meningitis			
HK1141	*H. influenzae*	NT	UK	III				ATCC19418 (NCTC4560)	
HK295	*H. influenzae*	NT	Egypt	III		Conjunctivitis			[71]
HK389	*H. influenzae*	NT		II				NCTC8143T	[71]
HK1212	*H. influenzae*	NT	Australia	III		BPF-like infection	Xylose negative	199/88	[7,10]
HK367	*H. aegyptius*	NT	USA	III		Conjunctivitis		NTCC8502	
HK1246	*H. aegyptius*	NT	USA	III		Conjunctivitis		Pittman 46	
HK1214	*H. aegyptius*	NT	Brazil, Valparaiso	III		Meningitis, CSF[b]	Plasmid negative	677/90	[9]
HK1240	*H. aegyptius*	NT	Brazil, Mato Grosso	III		Conjunctivitis	Plasmid restriction 1947	105/91	[9]
HK1219	*H. influenzae* biogr. aegyptius	NT	Brazil	III		Conjunctivitis, patient underwent intensive antibiotic treatment to prevent BPF	Plasmid negative, closely related to HK870 by MLST and MLEE	690/90	[9,42]
HK870	*H. influenzae* biogr. aegyptius	NT	Brazil	III		Septicemia	BPF case clone reference	F3031	[9]
HK2111	*H. haemolyticus*	NT	USA			COPD[c]	Hemolytic	3P5H	[72]

[a] NT, non-typeable.
[b] CSF, cerebral spinal fluid.
[c] COPD, chronic obstructive pulmonary disease.

Locus Enzyme Electrophoresis (MLEE) and MLST [7] (Fig. 1). These strains represent a broad phylogenetic, phenotypic, and geographic coverage of the species (Table 1). The bacterial strains here referred to as *H. influenzae*, *H. influenzae* biogroup aegyptius (Hibae), and *H. aegyptius* (Hae) all belong to one species according to traditional taxonomic criteria. The formal obstacle to merging all into one species *H. influenzae* is that the name *H. aegyptius* has priority. Until this problem is formally resolved we use the separate names to indicate that, although genetically closely related, the three taxa represent distinct populations of bacteria with distinct pathogenic potential. Genomic DNAs (gDNA) were isolated as described [7]. Metabolic profiling of the strains was performed as described by Kilian [1] and Brenner et al. [9].

2.2. Gene Discovery: HK1212 genomic library preparation and screening

The Gene Discovery approach is summarized in Fig. 1s. Briefly, a partial *Tsp*509I digestion was performed with 3 μg HK1212 to generate a mass peak centered at ~500 base pair fragments. Digested DNA was electrophoresed on 1% ultra pure agarose gel (Invitrogen, Carlsbad, CA), and fragments of an apparent MW ~300–800 bp were excised and recovered using the QIAquick gel extraction kit as per manufacturer's instruction (QIAgen, Valencia, CA). These fragments were cloned into the *Eco*RI site of pUC19 [20] and transformed into ElectroMAX DH10B cells (Invitrogen, Carlsbad, CA) by electroporation. High-throughput plasmid purification was performed at the J. Craig Venter Science Foundation, Joint Technology Center (Rockville, MD). To estimate the number of plasmids to screen, we assumed an average genomic insert size to be 500 bp and wished to set a 99% confidence limit of achieving 5× coverage of the HK1212 genome (1.8 Mb). Using the following equation: $N = \ln(1-P)/\ln(1-f)$, where: P = probability of cloning a given sequence, f = (average insert size)/(genome size), and ln = natural log [21], indicated that the screening of 18,000 recombinant plasmids were required. Plasmids were printed and screened using HK1212 and KW20/Rd genomic probes labeled with either Cy3 or Cy5. Replicate hybridizations were conducted with flip-dye replicates. The signal intensity ratios were used to indicate those plasmids that contained inserts that were significantly divergent or uniquely present in the HK1212 genome relative to KW20/Rd. Genomic DNA hybridizations were performed essentially as described at http://pfgrc.jcvi.org/index.php/microarray/protocols.html [22]. The slides were dried and then scanned using the GenePiX 4000B scanner (Axon Instruments, Union City, CA).

2.3. HK1212 genomic DNA sequence analysis

All sequence reads from selected recombinant plasmids (see previous section) were assembled using the TIGR assembler [23]. The resulting contigs and singletons were concatenated into a pseudomolecule [17,24]

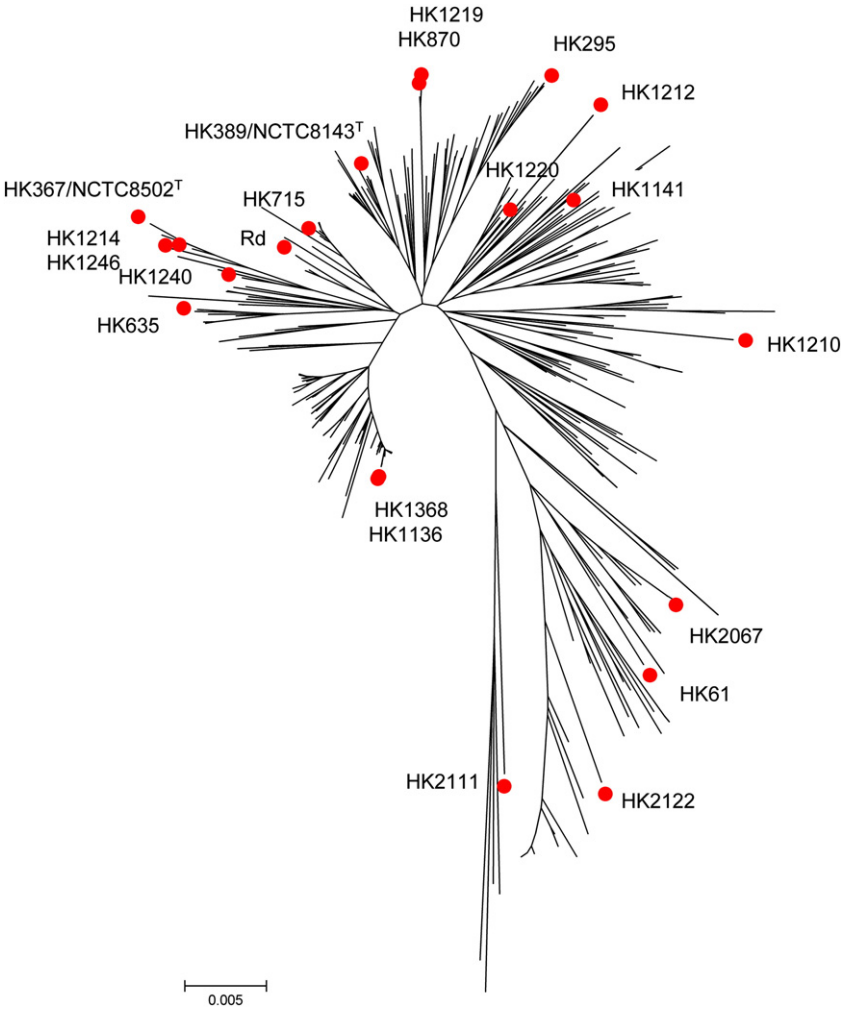

Fig. 1. MLST-based phylogenetic tree using *H. influenzae* sequences available in the MLST database (www.mlst.net). Red dots denote strains used in this study. Relationships are drawn using the minimum evolution algorithm applied to concatenated sequences of the seven MLST housekeeping genes and a fragment of *hap* [7]. The MLST-based calculated mean genetic distance within our collection is 0.026 ± 0.002, which is in agreement with the one calculated from all recognized sequence types (STs), i.e., 0.025 ± 0.002.

and subjected to the JCVI's automatic annotation pipeline, followed by manual curation. Glimmer [25] was used to identify open reading frames (ORFs). Proteins were also classified into putative functional role categories within the cell according to modified Riley's schema [26]. HK1212 genomic sequences obtained were initially filtered for uniqueness using the blastn algorithm [27] against KW20/Rd. Sequences with nucleotide matches <100 nucleotides and *P*-values $>10^{-5}$ were considered to be potentially unique. The sequences were further subjected to blastp [27] comparison against the non-redundant amino acid (NRAA) database. Sequences with a translated *P*-value match of $<10^{-5}$ were considered orthologous to previously characterized proteins while the remaining template sequences, i.e. those with no match or *P*-values $>10^{-5}$, were considered unique ORFs. In addition, we conducted gene cluster analysis to identify sequence variants that may represent paralogous (multi-gene) families using CAP3 [28]; sequences with <94% identity in the overlap region of >30 bp were considered paralogous genes. Sequence alignments were also investigated through Clustal X [29].

2.4. DNA microarray design and fabrication

ArrayOligoSelector (http://arrayoligosel.sourceforge.net [30]) was used to design 4578 seventy base oligonucleotides (70-mer genomic markers) for the *H. influenzae* multi-genome microarray. These oligonucleotides represent 4578 putative ORFs identified from complete and partial genomic sequences of *H. influenzae* KW20/Rd [14], 86-028NP [15], R2846 (NCBI accession AADO00000000), R2866 (NCBI accession AADP00000000), HK1212, *H. influenzae* biogroup aegyptius genetic island [31], and plasmids pF3031 and pF3028 [13]. Table 1s summarizes the 70-mers according to the number of features they represent from each sequence source. An equal volume of the oligonucleotides 50 μM (Illumina, San Diego, CA) were combined with 100% DMSO or Corning Buffer (Corning, NY). Oligonucleotides were printed onto Ultra-gap glass slides (Corning, NY) using a Lucidea microarray printer (Amersham, Piscataway, NJ). Printed slides were UV-crosslinked using a Spectrolinker XL-1500UV cross-linker (Spectronics Corporation, Westbury, NY) at 25,000 μJ/cm^2. Gene Discovery slides were generated by printing plasmid template DNA at a concentration of ~200 ng/μl.

2.5. CGH data analysis

Microarray data were analyzed as previously described [22]. Briefly, raw microarray images were analyzed using SpotFinder version 2.2.2 of TM4, Microarray Software Suite [32]. Spot florescence intensities corresponding to Cy3 and Cy5 were normalized using a log mode centering; spots less than 20,000 relative fluorescence units were filtered out from further analysis. The Histogram Mode Centering (HMC) algorithm [33] was used initially to normalize the signal data and fit the log-ratio frequencies from the main peak to a Gaussian distribution function. We utilized an analytical approach by estimating the best Gaussian fit through minimizing the sum of squared residuals between raw histogram data and a theoretical Gaussian curve. The standard deviation value that resulted from the model was used to calculate the percentiles corresponding to each log-ratio based on the z-score table. These calculated percentile values from the *z*-score table were used later to infer gene presence or absence in both reference and query genomes. We used conservative confidence limits to assign each gene represented in the array as present or absent in the reference and/or query genomes. These criteria have been visually summarized in Fig. 2s. Briefly, a gene was considered present with no significant sequence divergence, when the signal log-ratio of the corresponding 70-mer fell in the range of 5th–95th percentiles of the Gaussian fit model used for the normalization (see above). This group of genes was designated as shared (SH). Because we always calculated signal ratios as reference/query, a coding DNA sequence was considered uniquely present in the query (absent in the reference), if its 70-mer log-ratio percentile was ≤2.5. Alternatively, when the log-ratio percentile for a 70-mer was ≥97.5, the gene represented by that 70-mer was considered uniquely present in the reference (absent in the query). Areas within percentile values 2.5–5 or 95–97.5 were considered as borderline statistical confidence limits. Based on the hybridization signals, these two groups of features are considered to have some degree of sequence variability tending toward either the query or reference genome. Therefore, ORFs whose calculated 70-mer log-ratio percentiles fell in ranges of 2.5–5 or 95–97.5 were designated as divergent. Despite some noticeable sequence variability trends, ORFs classified into these two groups were considered present in both the reference and query genomes.

2.6. CGH data clustering

In order to minimize bias and noise due to the individual log-ratio values when investigating the global gene presence/absence patterns among the query strains, the final data set were assigned one of three different values: 0 = absent, 0.5 = divergent and 1 = present. A similar approach has been previously reported [34–36]. Hierarchical clustering of the query genomes based on the CGH data was conducted on both the Multiple Experiment Viewer (MeV) of TM4 Software and MrBayes as previously described [32,36–39]. Due to the type of the data, we applied non-parametric statistical analysis using Pearson Correlation algorithm from the MeV. MrBayes software, which has been developed for the Bayesian estimation of phylogeny [38], was used to further refine phylogenetic inferences.

2.7. Allelic grouping and gene designations

Our 70-mer design allowed us to distinguish subtle sequence differences among many sequence variants of orthologous and paralogous genes. In order to avoid confounding the final gene calls (presence, absence, and divergence), or the total gene complement estimates, we identified and clustered related orthologous sequence variants into “allelic groups.” Each allelic group, to the best of our ability, includes putative sequence variants of a group of orthologous coding DNA sequences (CDSs). Clustering of the orthologous sequences into allelic groups was based primarily on 70-mer-to-gene and protein-to-protein relationships. Allelic grouping was further refined by taking into consideration predicted features' functional roles as well as their corresponding Clusters of Orthologous Groups (COGs) assignments. In order to avoid under-estimating the total gene counts, allelic groups having one or more paralogous gene members were sub-divided so that each paralogous variant had its own sub-group when unambiguous data permitted.

2.8. Statistical approach for marker association analysis

We used Fisher's exact test for conducting marker association analysis (MAA) between the genomic attributes i.e. gene presence or absence, and groups of strains that possessed particular phenotypes or clades [40,41]. Markers, and consequently their corresponding ORFs, were considered “characteristic for,” “associated with,” or “prevalent in” a particular group inferred by MAA using $p<0.05$ as a cut-off value for statistical significance.

3. Results

We wished to establish the genetic basis for the emergence of *H. influenzae* causing BPF (Table 1). We initially screened 20 strains by CGH using a single-genome-based DNA microarray and conducted a phylogenomic cluster analysis (Fig. 3s). Based on that cluster analysis, strain HK1212 appeared the closest to the node of a divergent clade comprising all three BPF-associated isolates, as well as all Hae strains. This finding suggested that HK1212 represented one of the deepest

branching of the BPF clade. All three BPF isolates i.e. HK1212, HK870 and HK1219 displayed very similar gene loss patterns with respect to KW20/Rd. Strains HK870 and HK1219 had been isolated during the BPF outbreak in Brazil. Strain HK1212 was isolated in central Australia in 1986 from a child with characteristic symptomology of Brazilian Purpuric Fever (BPF, [5,10]). We hypothesized that the HK1212 genome contains additional or unique genes that contribute to its virulence, including its ability to evade innate immune factors, despite lacking a capsule.

Initially we screened a genomic library derived from HK1212 to identify novel sequences encoded in its genome. The sequencing of unique DNA segments resulted in the identification of a large number of HK1212 ORFs that were then combined with existing DNA sequence available in public databases allowing the design of a *H. influenzae* "species" DNA microarray. We utilized this array to screen a diverse set of Hi strains for their genomic content. The data generated from this comparative genomic hybridization approach allowed us to conduct a phylogenomic analysis for understanding the evolutionary premises and steps that are associated with the emergence of BPF clones.

3.1. Identification of unique genomic regions in strain HK1212, a BPF-like isolate

We developed a directed sequencing strategy we refer to as Gene Discovery (GD) that, like subtractive hybridization, enables the identification and rapid characterization of strain-specific sequences present in microbial genomes. The method (Supplementary Fig. 1s) exploits the use of DNA microarrays to discriminate between HK1212 unique and divergent genomic fragments to those shared by HK1212 and KW20/Rd genomes.

A total of 18,000 recombinant plasmids derived from an HK1212 genomic library were printed onto glass slide DNA microarrays. The genomic library was intentionally biased toward small DNA inserts to allow the identification of unique fragments whose length did not exceed the average length of eubacterial ORFs. DNA microarray hybridizations were conducted in replicate using flip-dye labeled genomic DNA probes (HK1212 and KW20/Rd). Hybridization signals displaying a HK1212/KW20/Rd $\log_2$-ratio of 2.0 or greater were identified and the 3309 corresponding plasmid DNAs were subjected to DNA Sanger sequencing via two end reads using vector-based primers.

The sequence reads were initially compared to the KW20/Rd genome to identify unique genomic sequences (i.e. present in HK1212 but absent in KW20/Rd). Out of the 6618 reads that were obtained from paired-end sequencing of selected plasmids, 3334 were considered unique. The remaining 3284 reads represented orthologous sequences present in both genomes with varying degrees of nucleotide divergence. Fig. 2A summarizes the results of the blastx searches of the non-redundant amino acid (NRAA) database.

All sequence reads were then assembled using TIGR assembler [23] resulting in 1302 contigs and 70 singletons. Contig sizes ranged from 138 to 3740 bp in length, with an average coverage of 3.2×. Annotation of the contigs produced 1459 putative ORFs or partial ORFs (pORFs), which are also referred to as "features." Supplementary Table 2s summarizes the characteristics of all novel HK1212 features discovered. Among the 1459 features, 439 were novel in HK1212 genome with respect to the *H. influenzae* KW20/Rd genome while 217 were unique i.e. no match to any hitherto sequenced genomes. The vast majority (90%) of the remaining novel features found in HK1212 reflected best hits to sequences from *Haemophilus* spp. and other members of the *Pasteurellaceae* family (Fig. 2B).

Analysis of the predicted functional role categories of those sequences unique to strain HK1212 genome relative to KW20/Rd is shown in Fig. 3. This comparison revealed significantly differential profiles between missing and unique gene functions encoded in the HK1212 relative to KW20/Rd. The group of genes encoding proteins of unknown functions (hypothetical proteins) constituted the largest gene set among HK1212 unique ORFs (38%); six out of 253 hypothetical proteins had matches in the EMBL's phage sequence database. Among the remaining groups of features with predicted functions we noticed that missing gene functions in HK1212 relative to KW20/Rd and their corresponding cellular role categories were not equally compensated by gained gene functions. While ORFs predicted to be involved in cellular transport, and energy metabolism constituted the majority of missing functions, those encoding mobile elements, cell envelope proteins, DNA metabolism, protein synthesis, protein fate, and amino acid biosynthesis made up the majority of HK1212 novel genomic content. We also searched the KEGG database to identify metabolic pathways that may be incomplete in HK1212 i.e. affected by gene absence in its genome. We focused our analysis on pathways containing ≥ 3 genes, and conservatively considered a pathway to be incomplete if ≥ 2 genes were absent. Results of this survey indicated that over nine pathways may be incomplete in HK1212 relative to HK20/Rd affecting the metabolism of: sucrose, glyoxylate and dicarboxylate, pentose and glucuronate interconversions, ascorbate and aldarate, butanoate, pantothenate and CoA biosynthesis, glycine, serine and threonine, lysine degradation, and vitamin B6. Some of the affected pathways may have downstream effect on others with which they are linked. For example, besides the absence of all ABC transporter components involved in xylose uptake, HK1212 is missing three genes involved in intracellular xylose metabolism — two xylulose kinases (HI1112 and HI1027) and a xylose isomerase (HI1112); these three enzymes are primarily involved in pentose and glucuronate interconversion pathway. However, xylose metabolism is linked to several other biochemical pathways involved in the metabolism of pentose phosphate, riboflavin, ascorbate and aldarate, and some amino sugars and nucleotide sugars; one of these three genes — xylulose kinase (HI1112), is also involved in fructose and mannose metabolism.

3.2. Identification of virulence factors in H. influenzae HK1212

Among the 439 HK1212-specific features (relative to KW20/Rd), we identified 80 (18%) predicted to encode virulence factors (Table 2). Virulence factors that facilitate host–pathogen interaction in the genome of HK1212 are proteins that are involved in invasion and cytadherence. Among the other virulence factors in *H. influenzae* HK1212 that do not directly facilitate host–pathogen interaction we include those that are involved in iron acquisition immune evasion proteins that contribute to phase or antigenic variation, and those that contribute to maintenance of cell envelope stability. Interestingly, the acquired repertoire of putative virulence factors in the HK1212 genome was dominated by genes encoding proteins predicted to facilitate the host–pathogen interaction, i.e. cytadherence and invasion. We found several features with sequence identity to known *H. influenzae* and *Neisseria* spp. virulence factor sequences including two BPF invasins *bpf001* and *bpf002* [12], *iga1* protease [7,42], *Haemophilus* adhesion and penetration protein — *hap* [43], adherence-related high molecular weight proteins *hmwA* and *hmwC* [44], *las*/*lav* (virulence associated protein A (*vapA*-) homologs, see below), fimbriae and fimbrial ushers, precursors of hemoglobin/haptoglobin binding protein A (*hgbA*), and heme/hemopexin utilization protein [14,45]. Some of the features encoding fimbriae, fimbrial ushers and haemaglutinins were found in multiple contigs and showed significant sequence divergence from their homologs within the novel gene set suggesting that they may represent paralogous families, a finding that is consistent with previous reports [48].

In an attempt to identify additional virulence associated genes, we examined features with significantly disparate GC content, those containing tetrameric repeats (Table 2s), or with significant sequence

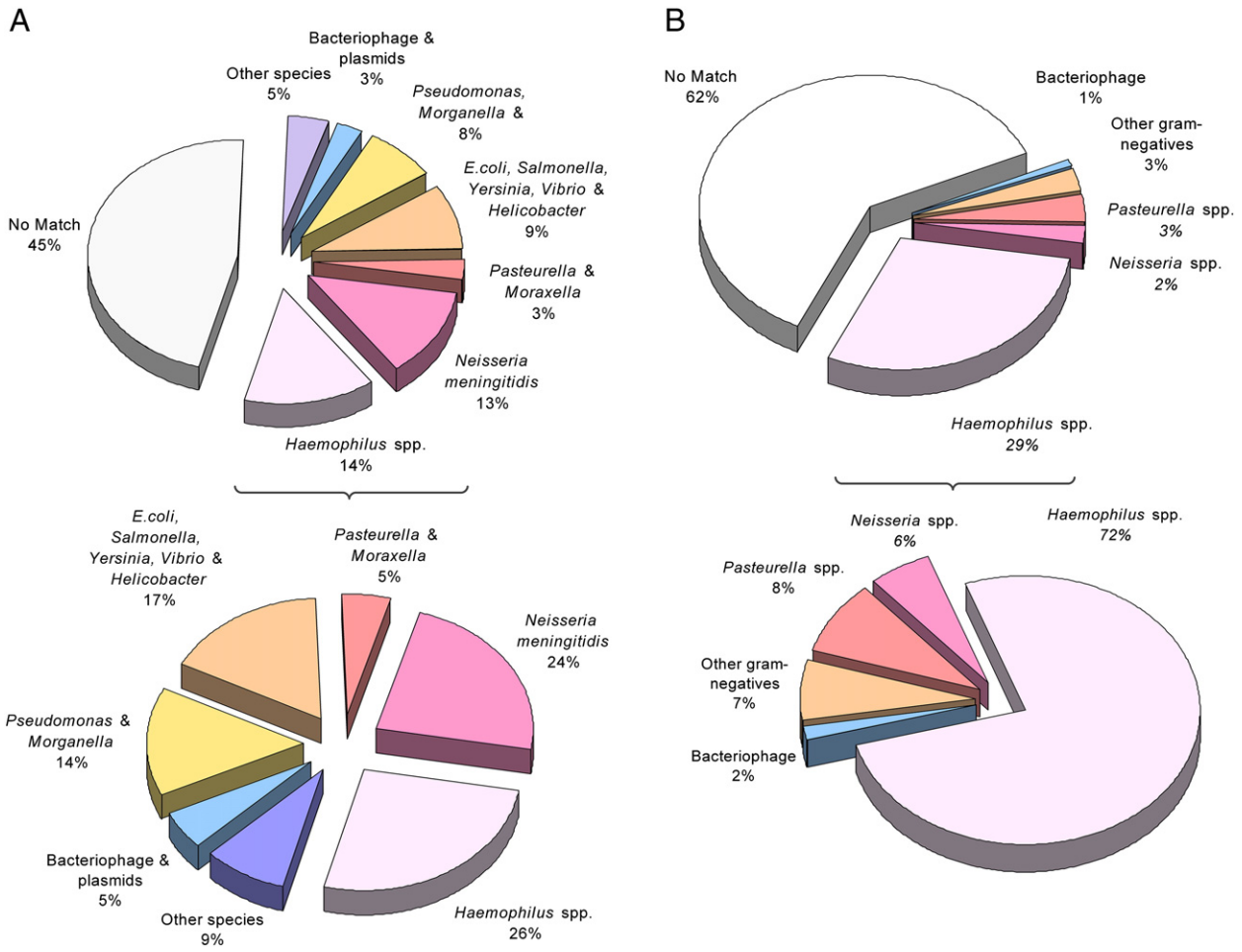

Fig. 2. Putative sources of the novel genomic content of HK1212 based on blast analysis. A. Summary of HK1212 sequence database searches using blastx. Species-associated sequences with the best blastx scores are shown with (top) and without (bottom) those with no match (hypotheticals). B. Same as A, except that the summary represents HK1212 features (partial ORF) analysis using blastp. Significant blast scores to sequences from non-typeable *H. influenzae* and *H. aegyptius* strains constitute the majority of "*Haemophilus* spp." fraction.

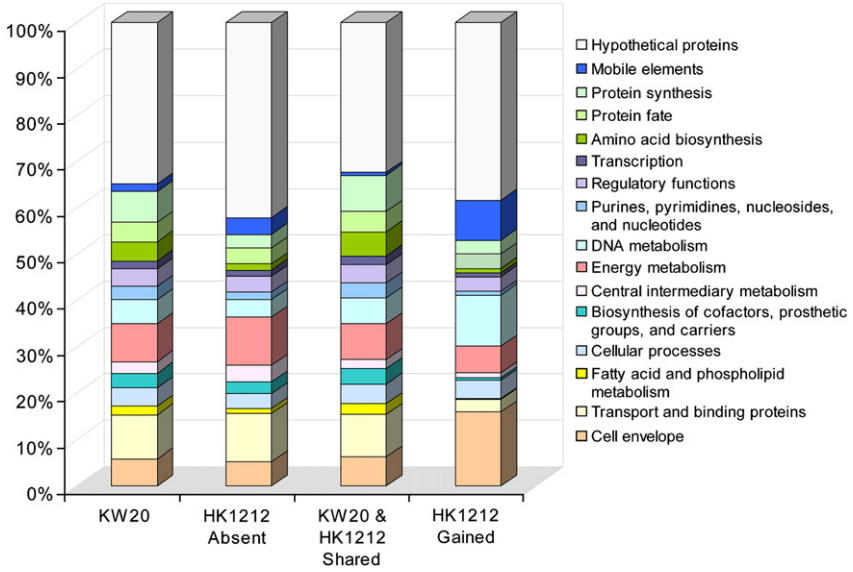

Fig. 3. Comparative sequence analysis of shared, presumed absent and gained cellular role categories in HK1212 relative to the *H. influenzae* KW20/Rd reference genome. Absent and shared role categories were determined by CGH using a single-genome array based on *H. influenzae* KW20/Rd. Gained functions were identified based on the Gene Discovery (GD) approach.

Table 2
Summary of novel features in HK1212 encoding putative virulence determinants.

Feature name	Minimum number of predicted paralogous groups
Adhesins (*emaA*, *xadA*)[A]	
Adhesins/autotransporters (*hsf*, *las*, *lav*, *pII*, *hep_hap*)[A]	
bpf001[V]	
bpf002[V]	
Fimbriae gene cluster *HifABCDE*[A]	2*
Other fimbriae and pili homologs[A]	2
Other fimbriae- and pili-usher homologs[A]	2
hmwA[A]	
hmwC[A]	
Haemaglutinins[A]	2
uspA1[A]	
licABC[‡,+,E]	
lex2AB[‡,+,E]	
Hgb[F]	
slpA[S,A]	
plp4[A]	

[A] Contributes to cytadherence.
[V] Contributes to invasion.
[F] Contributes to iron acquisition.
[S] Contributes to stability of membrane structures.
[*] Evidence for *hifB* only.
[‡] These gene loci are involved in LPS biosynthesis and degradation.
[+] One of the genes in this group contains tetrameric repeats that may contribute to phase or antigenic variation.
[E] Contributes to immune evasion.

identity to other members of *Pasteurellaceae* family or unrelated pathogens that share the same ecology with *H. influenzae*.

The average and the 95th percentile of the GC content within the KW20/Rd genome are 38% and 43.5%, respectively. We identified 50 (11%) novel features with GC content greater than 43.5%. The vast majority of them encode proteins involved in unknown functional roles (14, 28%), cell envelope (13, 26%), mobile elements (9, 18%), and DNA metabolism functions (6, 12%; Table 2s). Within this group of high-GC features, we also found 12 pORFs predicted to encode putative virulence factors, including putative pili and pili ushers, outer membrane proteins, and adhesins.

Blast analysis revealed several genes with a close relationship to genes resident in *N. meningitidis* including an *opaB* (pII-homolog, ORF00736) precursor and two *vapA* orthologs (ORF00319 and ORF00012). Opa proteins have been shown to be major virulence factors in both *N. meningitidis* and *N. gonorrhoeae* contributing to colonization and invasion [46–48]. Furthermore, *opa* genes encode highly variable surface exposed proteins that contribute to antigenic variation for evading the host immune response. *Neisseria* spp. vapA-homologs, *lav* and *las*, represent autotransporters found in *H. influenzae* and *H. influenzae* biogroup aegyptius, respectively [49]. The GC content of *N. meningitidis* serogroup B *lav* (*vapA*-homolog) is 40.0% closer to *H. influenzae* than *Neisseria* spp. (average of 52%). Therefore it has been hypothesized that this gene was a product of a recent transfer from *Haemophilus* to *Neisseria* [49–52]. Protein database searches revealed that ORF00109 and ORF00300 shared significant homology with the "Ubiquitous Surface exposed Protein A" (UspA1) from *Moraxella catarrhalis*. Although phylogenetically distant, this gram-negative aerobe is also a respiratory tract pathogen sharing a common environment with *H. influenzae*. *M. catarrhalis* UspA1 and its homolog, UspA2, are high molecular weight outer membrane proteins. It has been demonstrated that these proteins mediate cytadherence and biofilm formation [53–56].

Finally, we investigated the distribution pattern of HK1212 VAGs among 16 publicly available *H. influenzae* genomes. Results of this analysis indicated that less than half of the HK1212 VAG set (39 ORFs) have orthologs in other *H. influenzae* genomes; numbers vary between 15 and 35 ORFs/genome. The frequency of all these ORFs among the 16 genomes is also highly variable ranging 6–100%. Taken together, it is evident that the evolution of the strain HK1212 involved the acquisition of a comprehensive and large number of virulence factors via multiple horizontal gene transfer events. These events involved exchanges predominantly with its closest phylogenetic relatives and other pathogens residing in the same host environment.

3.3. H. influenzae multi-genome DNA microarray design

We performed bioinformatic analysis on publicly available sequences (NCBI release 152) from *Hi* genomes as well as those discovered from strain HK1212 in this study, to identify unique genes encoded in these genomes. The unique genomic features represented by 4578 unique 70-mer oligonucleotides in this array define approximately 2617 orthologous (alleic) groups, which were then used as the basis for estimating the genomic content of the query strains using CGH data.

In order to assess the comprehensiveness of the DNA microarray with respect to gene representation of the *H. influenzae* species and its member clades, we compared gene number estimates based on hybridization data for each of the 21 query genomes to genome size estimates based on PFGE. The error rate for the KW20/Rd gene count varied between 0.5% and 6% (average 3.7%) across all hybridization experiments. In addition, we found a strong linear correlation between the total gene estimates resulting from the CGH data to the genome size approximations determined by PFGE (Fig. 4). Taken collectively, our data indicate that the gene content represented on this microarray reflects a large proportion of the gene pool shared by Hi group members.

3.4. Phylogenomic relationships and clade-specific characteristics of Haemophilus strains

We further evaluated the phylogenomic relationships among *Haemophilus* group members with a focus on the Hi, HiBae and Hae, to identify genomic events associated with the BPF clone emergence. For this purpose, we have performed genome clustering using a variety of gene sets (Fig. 5). The first CGH dendrogram represents a summary of phylogenetic relationships based on global hybridization data wherein each gene is assigned the designation ("absent", "divergent" or "present"). The second dendrogram is based on a set of 329 markers corresponding to putative virulence associated genes, such as cytadherence-related proteins and iron transporters. The third dendrogram is based on the ratio percentiles of a set of 2097 conserved markers (present in ≥70% of the strains examined). Finally, the fourth dendrogram is based on 2481 markers corresponding to the variable genomic loci (present in <70% of the strains examined).

We noted an overall congruency among all dendrograms (Fig. 5). The dendrogram based on all genes indicated that the query strains form three major clades. The first one contained seven strains including all four Hae, two HiBae (Brazilian BPF) as well as the Australian BPF-like (HK1212). The second group consisted of five strains, two of which were from patients with septicemia, whereas the third group contained seven Hi, two originating from patients with meningitis. Regardless of the clustering approach (marker sets) used to investigate phylogenomic relationships, only Clade 1 (the "BPF/aegyptoids" clade) remained unchanged.

Genome clustering based on all genomic markers suggests that BPF clones are highly related to conjunctivitis isolates. As shown in Fig. 5, conjunctivitis isolates are located closer to the node of the clade compared to the BPF clones. Phylogenomic relationships among BPF/aegyptoids revealed through all CGH clustering approaches are consistent with previous studies based on phenotypic characterizations with regard to Hae and HiBae isolates [7].

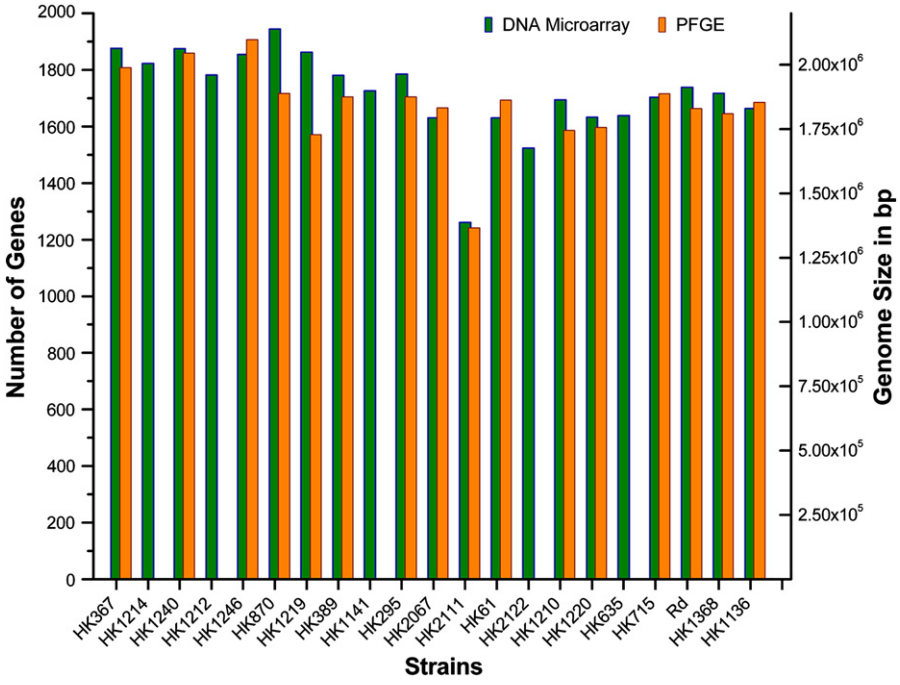

Fig. 4. Total gene content predictions derived from the CGH analysis using the species microarray and PFGE-based genome size estimates for the strains characterized in this study.

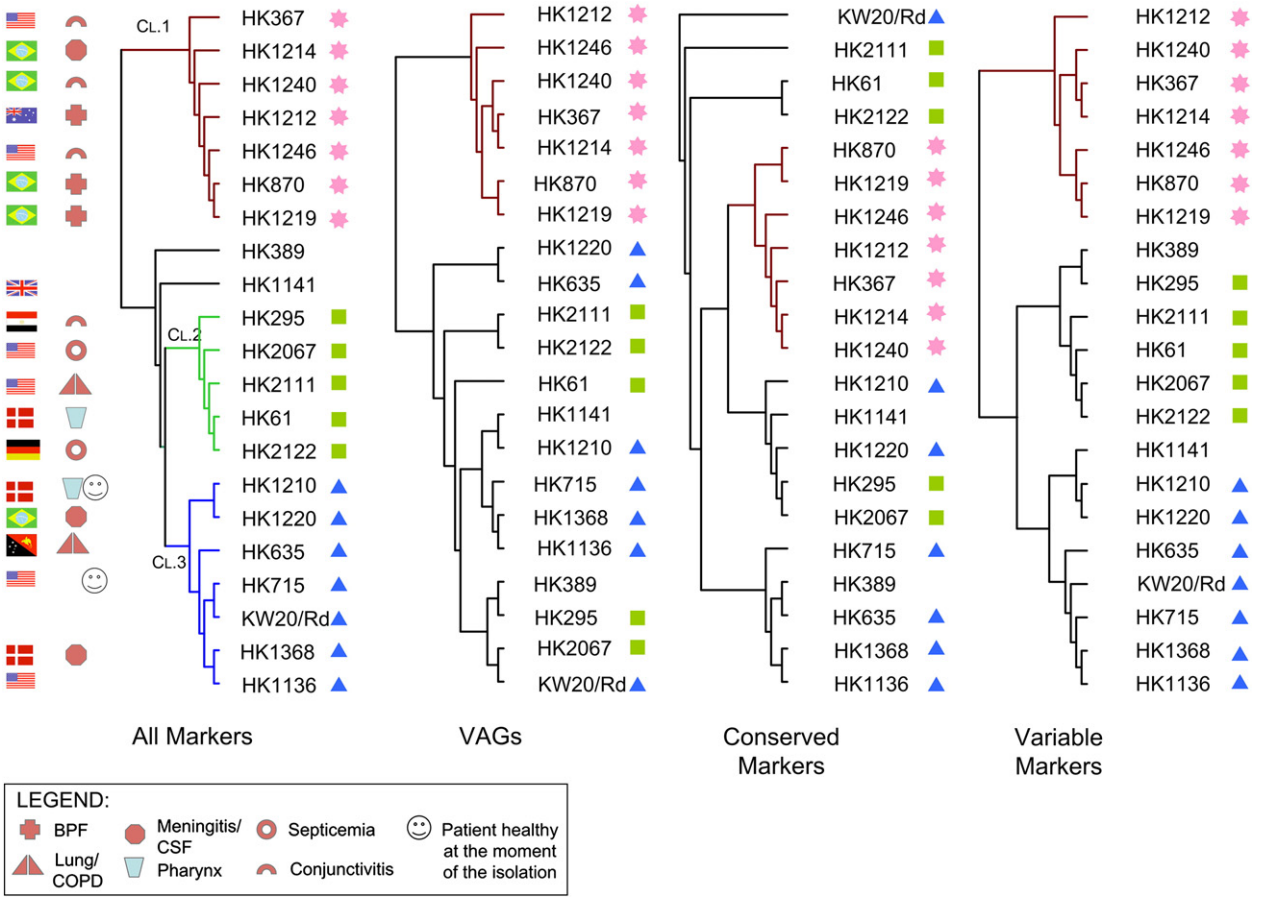

Fig. 5. Phylogenomic relationships among *Haemophilus* strains based on the CGH patterns using different markers sets. A. CGH clustering based on the global gene hybridization patterns of all 4578 70-mers using the trinary designations "0", "0.5" and "1" representing those coding sequences determined to be "absent", "divergent" and "present" respectively. Three major lineages, i.e. Clades 1, 2 and 3 are color-coded in brown, green and blue respectively. B. Dendrogram is based on the trinary designations from the 329 markers representing putative virulence associated genes. C. Clustering approach was based on the ratio percentile values of the conserved markers. D. Dendrogram is based on the trinary designations from the markers representing variable genes.

3.5. Cladistic analysis based on genomic flux

Based on the identified phylogenomic relationships, we conducted comparative analysis of the gene complements of isolates within each distinct clade. As illustrated in Fig. 4, BPF/aegyptoids (Clade 1) appeared to have somewhat larger genomes than the other Hi strains. We estimated that genomes belonging to BPF/aegyptoids possessed an average of 182 additional genes compared to all other clades (1859 vs. 1677 genes). Among BPF-related strains HK1212 appeared to share 83% of its gene content with HK870 and HK1219, while these latter two share 90–93% of their gene contents with each other.

In order to identify genomic events associated with the BPF clone emergence, we conducted a detailed marker association analysis (MAA) based on the Fisher's Exact Test. Initially we compared BPF/aegyptoids (Clade 1) with the remaining 14 query genomes. We identified 572 allelic groups that are characteristic of BPF/aegyptoids lineage (Table 3s). Among these clade-specific genomic events, 155 represent unique gene loss events, whereas 417 represent gene acquisition events ($p<0.05$), corresponding to a net gene gain of 262 genes (~280 kb). These genomic events resulted in significant functional modifications of strains within this lineage. Fig. 6 summarizes MAA results as they relate to predicted cellular functional role categories. Although the number of genes characteristic for the BPF/aegyptoids clade is relatively large, they correspond to a limited set of functions related most predominantly to mobile elements, proteins targeted to the cell surface, DNA metabolism and protein fate. Conversely, functions such as transport, energy metabolism, transcription, protein synthesis and fate, as well as synthesis of co-factors and amino acids were less prevalent among the BPF/aegyptoids. We also investigated the differential metabolic profiling between BPF/aegyptoids and the other *H. influenzae* strains in more detail. While we can identify apparently missing gene functions by CGH, we cannot rule out the possibility that the non-orthologous gene displacement events, including genes of unknown functions, have not occurred. We analyzed KEGG pathways containing ≥3 genes, and conservatively considered a pathway to be incomplete if ≥2 CDSs were absent in one clade compared to others. Results of this survey indicated that over 11 pathways may be incomplete or impaired due to gene absence among Clade 1 members. These functions impact the metabolism of: arginine and proline, glycine, serine and threonine, glyoxylate and dicarboxylate, pentose and glucuronate interconversions, fructose and mannose, amino sugars and nucleotide sugars, sucrose, glycerolipids, methane, one carbon pool by folate, and purines. Furthermore, BPF/aegyptoids had fewer proteins involved in transport functions than other strains — 19 vs. 26. Over eight transport systems [57] appear to be incomplete among BPF/aegyptoids with respect to other genomes — four ABC transporters, three ion channels and one secondary metabolite transporter. Besides the differences in the number of proteins involved in transports functions, we noticed a significant variation in the types of transporters that were characteristic for each group. For example, BFP/aegyptoids were devoid of the ABC transporter system responsible for xylose, cysteine, trace elements such as molybdenum, zinc and cobalt, while enriched for those involved in the uptake of galactosides, dipeptide/oligopeptide/nickel transport system, unspecified amino acids (COG0765) and thiamine. Although BPF/aegyptoids and other Hi strains appeared to have the same number of proteins involved in iron transport, there were still notable differences — *yfeD* (for chelated iron uptake, COG1108) type was characteristic for BPF/aegyptoids whereas *afuC* (for ferric uptake, COG3841) was characteristic for the other *H. influenzae* strains. The two-component regulatory system for sensing low levels of nitrogen from the environment also appears absent in BPF/

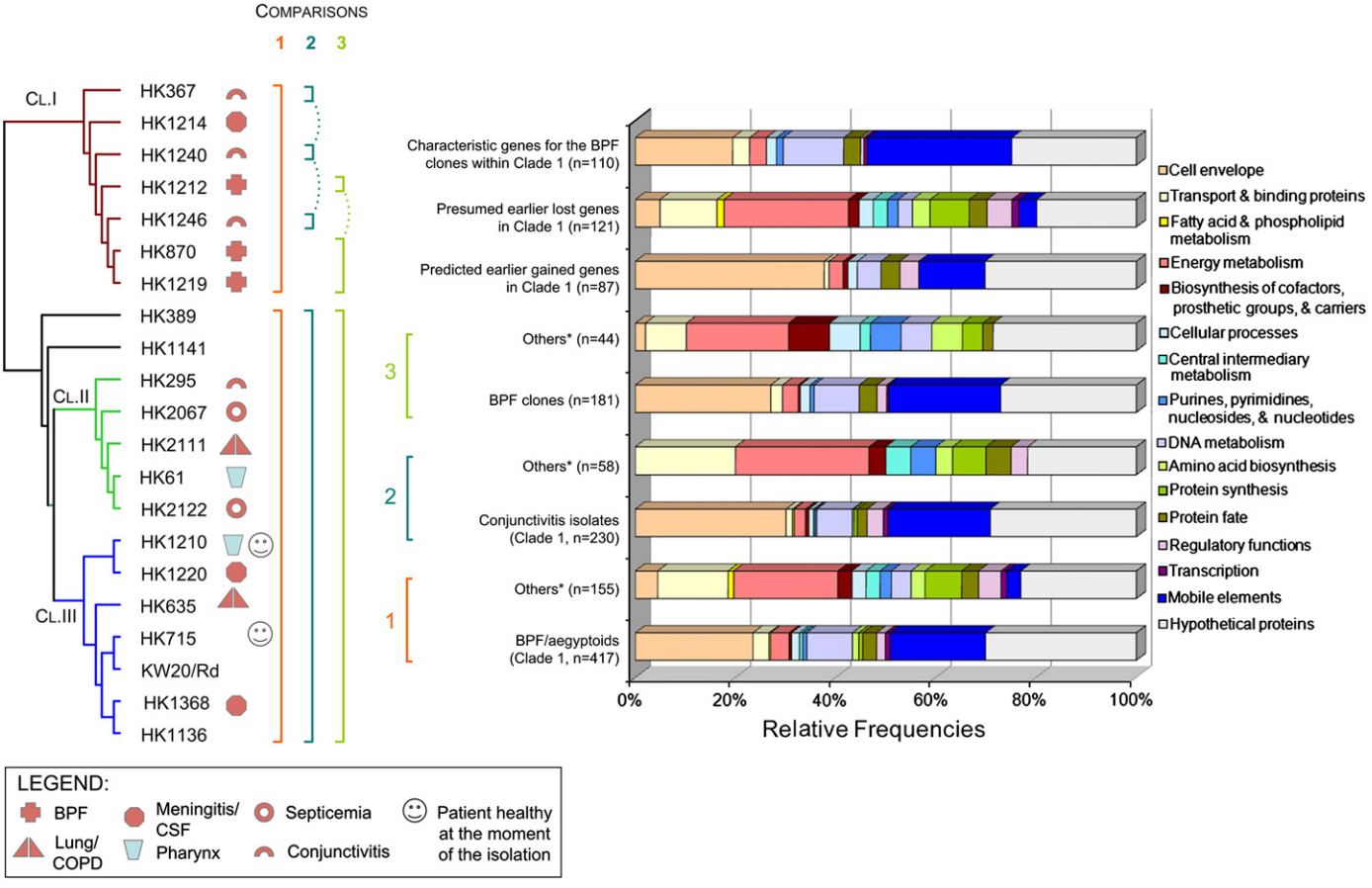

Fig. 6. Summary of functions characteristic for the BPF/aegyptoids (Clade 1) group members based on the marker association analysis using Fisher's exact test. Findings are summarized according to the putative cellular functional role categories of the coding DNA sequences. Numbers in brackets represent the number of characteristic genes for each group. The group of strains designated as "Others*" refer to Clade 2, Clade 3, HK369 and HK1141 genomes.

aegyptoids due to the absence of *glnD*. Finally, another gene of interest was tryptophan synthase beta subunit (*trpB*, HI1431) whose absence among all BPF/aegyptoids correlated 100% with the phenotypic assay for the tryptophanase activity [58].

After identifying characteristic features for the Clade 1 members (BPF/aegyptoids), we further applied the MAA within this group to delineate the extent of genome relatedness between conjunctivitis isolates and the BPF clones. Accordingly, we compared the genomes of three conjunctivitis isolates (HK367, HK1240 and HK1246) with the remainder of the Hi strains excluding BPF-associated isolates (HK1212, HK870 and HK1219). Next, we compared genomes of the BPF-associated isolates with the remainder of the Hi strains (including or excluding three conjunctivitis isolates of Clade 1). The direct comparison between the two sub-groups of Clade 1 was not conducted due to small number of genomes comprising each group. MAA results from these comparisons, summarized in the form of functional role categories, are shown in Fig. 6. Relative to the other Hi strains, we identified 230 and 181 CDSs (allele groups) characteristic for the conjunctivitis strains and the BPF-related isolates respectively ($p<0.05$). These two sub-groups shared 87 allele groups in common relative to the remainder of Hi clades (Table 3s). It is also worth noting here that HK1212-derived oligonucleotides that were added onto the multi-genome arrays constitute a large fraction of marker sets that were found characteristic for BPF/aegyptoids (74%), non-invasive conjunctivitis strains of Clade 1 (76%) and especially BPF case-related strains (89%).

In addition to identifying the shared genomic features that appeared early in the BPF/aegyptoids (Clade 1) evolution, we were able to identify those genes that are characteristic for either non-invasive conjunctivitis isolates or BPF-associated isolates. We found 110 CDSs that are characteristic for the BPF-related isolates. This gene set highlights the acquisition events associated with the emergence of the BPF clones. The majority (81%) of the gained BPF-specific novel functions were those related to cell envelope (18%), DNA metabolism (11%), mobile elements (27%) and genes with yet unknown cellular roles (24%). Among the 143 genes that are characteristic of Clade 1 non-invasive conjunctivitis strains we note the same trends in over-representation of functional role categories as the BPF clones. Eighty-five percent of CDSs specific for the conjunctivitis strains encoded proteins predicted to be localized in the cell envelope (25%), mobile elements (24%), DNA metabolism (8%), or proteins of unknown functions (28%).

We applied additional attention to the CDSs encoding putative VAGs to gain insights to the virulence of the BPF clones. The earlier acquisition events appear to have included sequence variants of genes encoding proteins involved in cytadherence, e.g. *hifB*, *hifC*, *Hsf*, *uspA1*, *hmwA*, *hep_hag* adhesins, pilins and pilin ushers (chaperones; Table 3s). Several of these genes had been previously found in HK1212 (Table 2). Interestingly, later events that may have taken place specifically in the genomes of BPF-related isolates involved the acquisition of additional genes functioning in cytadherence and invasion including *bpf001*, *bpf002*, *hifD*, a variant of *hifB*, heme/hemopexin-binding protein, and several pili and pili ushers unrelated to the previous group. Among the 181 genes found characteristic for the BPF-related strains only 38 (21%) are thought to contribute to the virulence. The CGH analysis also shows that the VAG set shared among BPF-related genomes does not have a uniform distribution among clades. For example, strains HK870, HK1219 and Clade 1 members as group share 84%, 81% and 47% of their VAG sets with HK1212 respectively. By contrast, the Clade 2 and Clade 3 group members share <1% of the HK1212 VAGs on the array.

Finally, Fig. 7 summarizes the MAA results for Clades 2 and 3 (Fig. 5), each containing two isolates associated with septicemia and two with meningitis, respectively. In contrast to the results regarding Clade 1 gene profiling, the number of absent CDSs exceeded the number of those found characteristic in Clades 2 and 3 by 2- and 4-fold respectively (Table 3s). It therefore appears that *H. influenzae* genome evolution is proceeding in such a way that some lineages are expanding their gene complements, while other clades may be moving toward more streamlined genomes. Our marker profile analysis indicated that each clade had its own particular sequence variants for genes involved in lipo-polysaccharide biosynthesis or outer membrane proteins (allelic groups 576, 745, and 1035). In addition, we noticed that members of the Clade 2 also seem to have characteristic sequence variants for the urease α-subunit and periplasmic-binding protein *yfeA*, which is involved in iron acquisition.

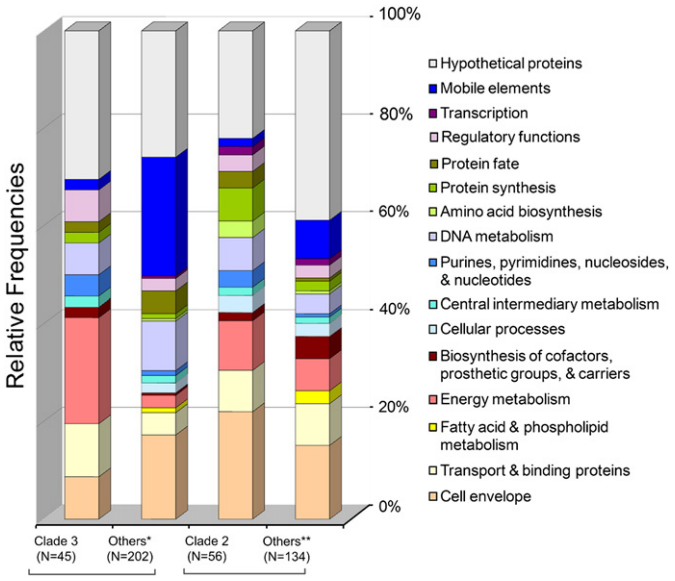

Fig. 7. Summary of functions characteristic for strains of Clades 2 and 3 (Fig. 3) based on the marker association analysis using Fisher's exact test. Findings are summarized according to the putative cellular functional role categories of the coding DNA sequences. Numbers in brackets represent the number of characteristic genes for each group. The group of strains designated as "Others*" refer to Clade 1, Clade 2, HK369 and HK1141 genomes, whereas "Others**" refers to Clade 1, Clade 3, HK369 and HK1141 genomes.

4. Discussion

This study using a *H. influenzae* species microarray enabled the elucidation of phylogenomic relationships among of members of this taxonomic group, in particular, clonal variants associated with BPF. In addition, by investigating the extent of *H. influenzae* species diversity, we identified numerous genomic features associated with the emergence of the BPF variants. However, the large number of genes of unknown function (~40%) discovered in the HK1212 accessory genome underscores that many gaps remain in our understanding of the genotype–phenotype relationship of these isolates. The genomic factors that contribute to emergence and the pathobiology of this highly virulent organism appear to be nominally related to a relatively large group of cell envelope proteins involved in cytadherence and invasion. Our findings support the previous hypothesis that BPF isolates have an unusual membrane composition, especially with regard to pili [44,59,60]. The features encoding for pili-related functions made up a significant portion of the putative virulence genes identified in the BPF-like isolate, HK1212. Mobile elements, almost exclusively composed of phage proteins, were also highly represented among acquired genes. Likewise, this group of features (mobile elements) represented ~20% of the genetic content characteristic for the BPF/aegyptoids genomes. This pattern is very similar to the one observed in the genome of enterohemorrhagic *Escherichia coli* O157:H7 [61], and this finding underscores the significance of mobile

elements as vehicles of horizontal gene transfer and as such, they contribute extensively to the generation of genetic diversity [62–64].

In conjunction with the acquisition of a limited set of acquired functional roles, it is apparent that genomic flux among BPF/aegyptoids genomes is characterized by significant alterations in selected role categories. Several genes involved in housekeeping cellular functions such as energy metabolism, transcription, translation, synthesis of co-factors, amino acids and proteins, as well as transport, appear to have been lost. Our findings suggest that, in addition to deficiencies in a few two-component regulatory systems and transport systems, several pathways involved in the metabolism of certain carbohydrates (xylose, sucrose, fructose, mannose and nucleotide sugars), vitamins, nucleotides and amino acid metabolism (arginine and proline, glycine, serine and threonine) are impaired in BPF/aegyptoids. Consequently, adaptation to the host environment may be driving the loss of genes that are no longer needed in the host environment. Without exception, the gene loss profile among Clade 1 members, and especially BPF-related strains, suggests that genome evolution is driving the bacterial cell toward deeper dependency on the host for energy and metabolites (e.g. loss of genes involved in energy metabolism and transport). This process has been accompanied by a sequential acquisition of genes associated with particular functional roles that have enabled the cell to invade host tissues with remarkable efficiency. Almost all VAGs found in HK1212 and characteristic for the BPF-related stains as a group appear to be involved in cytadherence and invasion.

Extensive research on *H. influenzae* pathogenesis has resulted in the identification and characterization of many virulence factors. However, the screening of clinical isolates has indicated that their distribution is not universal [65]. With the exception of the *cap* genes, we identified an unusually large set of previously described *H. influenzae* virulence factors in the genome of HK1212. Based on our literature review [15,16,19,65,66], CGH data, as well as comparative sequence analysis, this is quite different from known patterns of virulence gene content discovered within the genomes of various members of *H. influenzae*. Although the genome content revealed by our study is a snapshot of HK1212 evolution, it is clear that a gene complement supporting high invasiveness in a capsule independent manner has emerged. As a consequence, we postulate that during this developmental process, *cap* genes may have been counter-selected and lost. It is possible that the presence of the capsule may interfere or be incompatible with the function of other surface-localized virulence factors. This speculation is consistent with the demonstration that *cap*$^{-}$ pneumococci adhere more efficiently to epithelial cells and that such strains are more associated with conjunctivitis [67].

CGH-based phylogenomic analysis revealed three major groups, and for the majority of these strains, phylogenomic relationships were essentially congruent among all clustering approaches independent of the gene set used. The strongest relationships were observed among the seven members of Clade 1, including all four Hae, two HiBae (Brazilian BPF) as well as the Australian BPF-like (HK1212). Regardless of the clustering approach, the members of this group remained the same, indicating that "BPF/aegyptoids" form a coherent clade. The second clade included two strains obtained from patients with septicemia, whereas the third clade contained two isolates associated with meningitis. Despite major clade separation, it was noticed that virulent strains clustered together with those isolated from healthy individuals or patients with milder forms of disease.

All clustering approaches provided complementary information for elucidating overall genome content relatedness. The phylogenomic relationships inferred by clustering based on gene presence, in general, do not reproduce the phylogeny of a species in terms of vertical evolution, but instead represent the overall relatedness of genomes to one another, and provide a means to evaluate genome evolution within the species. The MLST clustering approach indicated initially that the two Brazilian BPF clones (HK870 and HK1219) are ancestrally distant from the Australian isolate (HK1212); these observations were further supported by the clustering based on conserved loci (Fig. 1). At the same time, we found that the BPF clones as well as their conjunctivitis close relatives share similar genomic compositions.

Hae and Hibae are recognized as a separate phylogenetic group within the species *H. influenzae* [1,7]. This group (containing both Hae and Hibae) is characterized by a distinct cell morphology, microcolony formation on conjunctival epithelial cells, a strong hemagglutinating activity, the inability to metabolize xylose, and a decaying *hap* gene [7]. Both Hae and Hibae cause acute eye infections but the latter has invasive potential similar to serotype b strains. It is remarkable that the phylogenetically distinct, non-encapsulated strains HK870, HK1219 and HK1212 share a unique and invasive pathogenic potential that is consistent with their patterns of gene acquisition and gene loss. Based on the *hap* pseudogene analysis, it has been hypothesized that the BPF-associated *H. influenzae* biogroup aegyptius and *H. aegyptius* evolved recently as a distinct group from *H. influenzae* [1,7]. The patterns of present and missing genes further suggest that the BPF-associated *H. influenzae* biogroup aegyptius (HK870 and HK1219) share a recent common ancestor with *H. aegyptius* (HK1246) and HK1212.

Based on the CGH results, this evolutionary hypothesis suggests that gene acquisition events resulting in the recent clone emergence, have taken place independently. Therefore, characterization of gene presence/absence patterns has fundamental epidemiological significance. This type of analytical approach will help improve diagnostics and open the avenues for developing predictive cladistic models to improve our understanding of genome evolution associated with clone emergence.

Future efforts will seek to broaden our coverage of genomes to include additional sequenced strains with the goal of identifying genomic markers that can be used for diagnostic purposes and finally assist both drug- and vaccine designs [24,68].

5. Sequence accession numbers

The sequence data from this study have been submitted to the NCBI under accession number ABFC00000000.

6. Microarray data deposition

The sequence data from this study have been submitted to the NCBI Gene Expression omnibus (GEO) under accession number GSE8300.

Acknowledgments

L. Papazisi and S. Ratnayake contributed equally to this work with writing and data analysis. B. Remortel and G. Bock contributed equally to this work with regard to the CGH data generation. We thank Dr. Marcus Jones and Dr. Timothy Minogue for their critical review of the manuscript, and Dr. Chun-Hua Wan for his assistance regarding microarray data formatting and submission to the GEO database. This work was supported by the NIAID contract No. N01-AI-15447 to the Pathogen Functional Genomics Resource Center (PFGRC) at the JCVI.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2010.07.005.

References

[1] M. Kilian, *Haemophilus*, in: P.R. Murray, E.J. Baron, J.H. Jorgensen, M.L. Landry, M.A. Pfaller (Eds.), Manual of Clinical Microbiology, 9th ed., American Society of Microbiology, Washinton D.C., 2007, pp. 636–648.
[2] E. Meats, E.J. Feil, S. Stringer, A.J. Cody, R. Goldstein, J.S. Kroll, T. Popovic, B.G. Spratt, Characterization of encapsulated and noncapsulated *Haemophilus*

influenzae and determination of phylogenetic relationships by multilocus sequence typing, J. Clin. Microbiol. 41 (2003) 1623–1636.
[3] H. Peltola, Worldwide *Haemophilus influenzae* type b disease at the beginning of the 21st century: global analysis of the disease burden 25 years after the use of the polysaccharide vaccine and a decade after the advent of conjugates, Clin. Microbiol. Rev. 13 (2000) 302–317.
[4] J.S. Kroll, B. Loynds, L.N. Brophy, E.R. Moxon, The bex locus in encapsulated *Haemophilus influenzae:* a chromosomal region involved in capsule polysaccharide export, Mol. Microbiol. 4 (1990) 1853–1862.
[5] L.H. Harrison, V. Simonsen, E.A. Waldman, Emergence and disappearance of a virulent clone of *Haemophilus influenzae* biogroup aegyptius, cause of Brazilian Purpuric Fever, Clin. Microbiol. Rev. 21 (2008) 594–605.
[6] J.M. Musser, R.K. Selander, Brazilian Purpuric Fever: evolutionary genetic relationships of the case clone of *Haemophilus influenzae* biogroup aegyptius to encapsulated strains of *Haemophilus influenzae*, J. Infect. Dis. 161 (1990) 130–133.
[7] M. Kilian, K. Poulsen, H. Lomholt, Evolution of the paralogous hap and iga genes in *Haemophilus influenzae*: evidence for a conserved hap pseudogene associated with microcolony formation in the recently diverged *Haemophilus aegyptius* and *H. influenzae* biogroup aegyptius, Mol. Microbiol. 46 (2002) 1367–1380.
[8] L.M. Smoot, D.D. Franke, G. McGillivary, L.A. Actis, Genomic analysis of the F3031 Brazilian Purpuric Fever clone of *Haemophilus influenzae* biogroup aegyptius by PCR-based subtractive hybridization, Infect. Immun. 70 (2002) 2694–2699.
[9] D.J. Brenner, L.W. Mayer, G.M. Carlone, L.H. Harrison, W.F. Bibb, M.C. Brandileone, F.O. Sottnek, K. Irino, M.W. Reeves, J.M. Swenson, et al., Biochemical, genetic, and epidemiologic characterization of *Haemophilus influenzae* biogroup aegyptius (*Haemophilus aegyptius*) strains associated with Brazilian Purpuric Fever, J. Clin. Microbiol. 26 (1988) 1524–1534.
[10] P. McIntyre, G. Wheaton, J. Erlich, D. Hansman, Brasilian Purpuric Fever in central Australia, Lancet 2 (1987) 112.
[11] S.R. Dobson, J.S. Kroll, E.R. Moxon, Insertion sequence IS1016 and absence of *Haemophilus* capsulation genes in the Brazilian Purpuric Fever clone of *Haemophilus influenzae* biogroup aegyptius, Infect. Immun. 60 (1992) 618–622.
[12] M.S. Li, J.L. Farrant, P.R. Langford, J.S. Kroll, Identification and characterization of genomic loci unique to the Brazilian Purpuric Fever clonal group of *H. influenzae* biogroup aegyptius: functionality explored using meningococcal homology, Mol. Microbiol. 47 (2003) 1101–1111.
[13] J.S. Kroll, J.L. Farrant, S. Tyler, M.B. Coulthart, P.R. Langford, Characterisation and genetic organisation of a 24-MDa plasmid from the Brazilian Purpuric Fever clone of *Haemophilus influenzae* biogroup aegyptius, Plasmid 48 (2002) 38–48.
[14] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, et al., Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, Science 269 (1995) 496–512.
[15] A.L. Erwin, K.L. Nelson, T. Mhlanga-Mutangadura, P.J. Bonthuis, J.L. Geelhood, G. Morlin, W.C. Unrath, J. Campos, D.W. Crook, M.M. Farley, F.W. Henderson, R.F. Jacobs, K. Muhlemann, S.W. Satola, L. van Alphen, M. Golomb, A.L. Smith, Characterization of genetic and phenotypic diversity of invasive nontypeable *Haemophilus influenzae*, Infect. Immun. 73 (2005) 5853–5863.
[16] A. Harrison, D.W. Dyer, A. Gillaspy, W.C. Ray, R. Mungur, M.B. Carson, H. Zhong, J. Gipson, M. Gipson, L.S. Johnson, L. Lewis, L.O. Bakaletz, R.S. Munson Jr., Genomic sequence of an otitis media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae* serotype d, strain KW20, J. Bacteriol. 187 (2005) 4627–4636.
[17] J.S. Hogg, F.Z. Hu, B. Janto, R. Boissy, J. Hayes, R. Keefe, J.C. Post, G.D. Ehrlich, Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains, Genome Biol. 8 (2007) R103.
[18] M.M. Fernaays, A.J. Lesse, S. Sethi, X. Cai, T.F. Murphy, Differential genome contents of nontypeable *Haemophilus influenzae* strains from adults with chronic obstructive pulmonary disease, Infect. Immun. 74 (2006) 3366–3374.
[19] K. Shen, P. Antalis, J. Gladitz, S. Sayeed, A. Ahmed, S. Yu, J. Hayes, S. Johnson, B. Dice, R. Dopico, R. Keefe, B. Janto, W. Chong, J. Goodwin, R.M. Wadowsky, G. Erdos, J.C. Post, G.D. Ehrlich, F.Z. Hu, Identification, distribution, and expression of novel genes in 10 clinical isolates of nontypeable *Haemophilus influenzae*, Infect. Immun. 73 (2005) 3479–3491.
[20] C. Yanisch-Perron, J. Vieira, J. Messing, Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors, Gene 33 (1985) 103–119.
[21] D.D. Moore, Overview of recombinant DNA libraries, in: F.M. Ausubel, R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, et al., (Eds.), Current Protocols in Molecular Biology, John Wiley & Sons, New York, 1993, pp. 5.1.1–5.1.3.
[22] S.N. Peterson, C.K. Sung, R. Cline, B.V. Desai, E.C. Snesrud, P. Luo, J. Walling, H. Li, M. Mintz, G. Tsegaye, P.C. Burr, Y. Do, S. Ahn, J. Gilbert, R.D. Fleischmann, D.A. Morrison, Identification of competence pheromone responsive genes in *Streptococcus pneumoniae* by use of DNA microarrays, Mol. Microbiol. 51 (2004) 1051–1070.
[23] G. Sutton, O. White, M.D. Adams, A.R. Kerlavage, A new tool for assembling large shotgun sequencing projects, Genome Sci. Technol. 1 (1995) 9–19.
[24] H. Tettelin, V. Masignani, M.J. Cieslewicz, C. Donati, D. Medini, N.L. Ward, S.V. Angiuoli, J. Crabtree, A.L. Jones, A.S. Durkin, R.T. Deboy, T.M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J.D. Peterson, C.R. Hauser, J.P. Sundaram, W.C. Nelson, R. Madupu, L.M. Brinkac, R.J. Dodson, M.J. Rosovitz, S.A. Sullivan, S.C. Daugherty, D.H. Haft, J. Selengut, M.L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K.J. O'Connor, S. Smith, T.R. Utterback, O. White, C.E. Rubens, G. Grandi, L.C. Madoff, D.L. Kasper, J.L. Telford, M.R. Wessels, R. Rappuoli, C.M. Fraser, Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome", Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 13950–13955.
[25] A.L. Delcher, D. Harmon, S. Kasif, O. White, S.L. Salzberg, Improved microbial gene identification with GLIMMER, Nucleic Acids Res. 27 (1999) 4636–4641.
[26] M. Riley, Functions of the gene products of *Escherichia coli*, Microbiol. Rev. 57 (1993) 862–952.
[27] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[28] X. Huang, A. Madan, CAP3: a DNA sequence assembly program, Genome Res. 9 (1999) 868–877.
[29] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.
[30] Z. Bozdech, J. Zhu, M.P. Joachimiak, F.E. Cohen, B. Pulliam, J.L. DeRisi, Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray, Genome Biol. 4 (2003) R9.
[31] G. McGillivary, A.P. Tomaras, E.R. Rhodes, L.A. Actis, Cloning and sequencing of a genomic island found in the Brazilian Purpuric Fever clone of *Haemophilus influenzae* biogroup aegyptius, Infect. Immun. 73 (2005) 1927–1938.
[32] A.I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, J. Quackenbush, TM4: a free, open-source system for microarray data management and analysis, Biotechniques 34 (2003) 374–378.
[33] C.C. Kim, E.A. Joyce, K. Chan, S. Falkow, Improved analytical methods for microarray-based genome-composition analysis, Genome Biol. 3 (2002) (RESEARCH0065).
[34] E.F. Boyd, S. Porwollik, F. Blackmer, M. McClelland, Differences in gene content among *Salmonella enterica* serovar typhi isolates, J. Clin. Microbiol. 41 (2003) 3823–3828.
[35] S. Porwollik, R.M. Wong, M. McClelland, Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 8956–8961.
[36] O.L. Champion, M.W. Gaunt, O. Gundogdu, A. Elmi, A.A. Witney, J. Hinds, N. Dorrell, B.W. Wren, Comparative phylogenomics of the food-borne pathogen *Campylobacter jejuni* reveals genetic markers predictive of infection source, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 16043–16048.
[37] H.L. Leavis, R.J. Willems, W.J. van Wamel, F.H. Schuren, M.P. Caspers, M.J. Bonten, Insertion sequence-driven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*, PLoS Pathog. 3 (2007) e7.
[38] F. Ronquist, J.P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics 19 (2003) 1572–1574.
[39] R.A. Stabler, D.N. Gerding, J.G. Songer, D. Drudy, J.S. Brazier, H.T. Trinh, A.A. Witney, J. Hinds, B.W. Wren, Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains, J. Bacteriol. 188 (2006) 7297–7305.
[40] J.H. Zar, Biostatistical Analysis, 4th ed.Prentice Hall, Upper Saddle River NJ, USA, 1999.
[41] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, 2004.
[42] H. Lomholt, M. Kilian, Distinct antigenic and genetic properties of the immunoglobulin A1 protease produced by *Haemophilus influenzae* biogroup aegyptius associated with Brazilian Purpuric Fever in Brazil, Infect. Immun. 63 (1995) 4389–4394.
[43] J.W. St Geme III, M.L. de la Morena, S. Falkow, A *Haemophilus influenzae* IgA protease-like protein promotes intimate interaction with human epithelial cells, Mol. Microbiol. 14 (1994) 217–233.
[44] J.R. Gilsdorf, C.F. Marrs, B. Foxman, *Haemophilus influenzae*: genetic variability and natural selection to identify virulence factors, Infect. Immun. 72 (2004) 2457–2461.
[45] H. Jin, Z. Ren, P.W. Whitby, D.J. Morton, T.L. Stull, Characterization of hgpA, a gene encoding a haemoglobin/haemoglobin–haptoglobin-binding protein of *Haemophilus influenzae*, Microbiology 145 (Pt 4) (1999) 905–914.
[46] M. Toleman, E. Aho, M. Virji, Expression of pathogen-like Opa adhesins in commensal *Neisseria*: genetic and functional analysis, Cell Microbiol. 3 (2001) 33–44.
[47] M.I. de Jonge, H.J. Hamstra, L. van Alphen, J. Dankert, P. van der Ley, Mapping the binding domains on meningococcal Opa proteins for CEACAM1 and CEA receptors, Mol. Microbiol. 50 (2003) 1005–1015.
[48] F.L. Naids, B. Belisle, N. Lee, R.F. Rest, Interactions of *Neisseria gonorrhoeae* with human neutrophils: studies with purified PII (Opa) outer membrane proteins and synthetic Opa peptides, Infect. Immun. 59 (1991) 4628–4635.
[49] J. Davis, A.L. Smith, W.R. Hughes, M. Golomb, Evolution of an autotransporter: domain shuffling and lateral transfer from pathogenic *Haemophilus* to *Neisseria*, J. Bacteriol. 183 (2001) 4626–4635.
[50] U.A. Ochsner, A.I. Vasil, Z. Johnson, M.L. Vasil, *Pseudomonas aeruginosa* fur overlaps with a gene encoding a novel outer membrane lipoprotein, OmlA, J. Bacteriol. 181 (1999) 1099–1109.
[51] S. Grizot, S.K. Buchanan, Structure of the OmpA-like domain of RmpM from *Neisseria meningitidis*, Mol. Microbiol. 51 (2004) 1027–1037.
[52] B.J. May, Q. Zhang, L.L. Li, M.L. Paustian, T.S. Whittam, V. Kapur, Complete genomic sequence of *Pasteurella multocida*, Pm70, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 3460–3465.
[53] D.J. Hill, M. Virji, A novel cell-binding mechanism of *Moraxella catarrhalis* ubiquitous surface protein UspA: specific targeting of the N-domain of carcinoembryonic antigen-related cell adhesion molecules by UspA1, Mol. Microbiol. 48 (2003) 117–129.
[54] E.R. Lafontaine, L.D. Cope, C. Aebi, J.L. Latimer, G.H. McCracken Jr., E.J. Hansen, The UspA1 protein and a second type of UspA2 protein mediate adherence of *Moraxella catarrhalis* to human epithelial cells in vitro, J. Bacteriol. 182 (2000) 1364–1373.

[55] M.M. Pearson, C.A. Laurence, S.E. Guinn, E.J. Hansen, Biofilm formation by *Moraxella catarrhalis* in vitro: roles of the UspA1 adhesin and the Hag hemagglutinin, Infect. Immun. 74 (2006) 1588–1596.
[56] E. Hoiczyk, A. Roggenkamp, M. Reichenbecher, A. Lupas, J. Heesemann, Structure and sequence analysis of *Yersinia* YadA and *Moraxella* UspAs reveal a novel class of adhesins, EMBO J. 19 (2000) 5989–5999.
[57] Q. Ren, K. Chen, I.T. Paulsen, TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels, Nucleic Acids Res. 35 (2007) D274–D279.
[58] K. Martin, G. Morlin, A. Smith, A. Nordyke, A. Eisenstark, M. Golomb, The tryptophanase gene cluster of *Haemophilus influenzae* type b: evidence for horizontal gene transfer, J. Bacteriol. 180 (1998) 107–118.
[59] T.D. Read, S.W. Satola, M.M. Farley, Nucleotide sequence analysis of hypervariable junctions of *Haemophilus influenzae* pilus gene clusters, Infect. Immun. 68 (2000) 6896–6902.
[60] T.D. Read, M. Dowdell, S.W. Satola, M.M. Farley, Duplication of pilus gene complexes of *Haemophilus influenzae* biogroup aegyptius, J. Bacteriol. 178 (1996) 6564–6570.
[61] T. Hayashi, K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, H. Shinagawa, Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12, DNA Res. 8 (2001) 11–22.
[62] L.S. Frost, R. Leplae, A.O. Summers, A. Toussaint, Mobile genetic elements: the agents of open source evolution, Nat. Rev. Microbiol. 3 (2005) 722–732.
[63] U. Dobrindt, B. Hochhut, U. Hentschel, J. Hacker, Genomic islands in pathogenic and environmental microorganisms, Nat. Rev. Microbiol. 2 (2004) 414–424.
[64] C.R. Woese, On the evolution of cells, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 8742–8747.
[65] N.J. High, *Haemophilus influenzae*, Medical Molecular Microbiolgy, Academic Press, 2001, pp. 1967–1988.
[66] R.S. Munson Jr., A. Harrison, A. Gillaspy, W.C. Ray, M. Carson, D. Armbruster, J. Gipson, M. Gipson, L. Johnson, L. Lewis, D.W. Dyer, L.O. Bakaletz, Partial analysis of the genomes of two nontypeable *Haemophilus influenzae* otitis media isolates, Infect. Immun. 72 (2004) 3002–3010.
[67] M. Martin, J.H. Turco, M.E. Zegans, R.R. Facklam, S. Sodha, J.A. Elliott, J.H. Pryor, B. Beall, D.D. Erdman, Y.Y. Baumgartner, P.A. Sanchez, J.D. Schwartzman, J. Montero, A. Schuchat, C.G. Whitney, An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae*, N. Engl. J. Med. 348 (2003) 1112–1121.
[68] T.D. Read, S.N. Peterson, N. Tourasse, L.W. Baillie, I.T. Paulsen, K.E. Nelson, H. Tettelin, D.E. Fouts, J.A. Eisen, S.R. Gill, E.K. Holtzapple, O.A. Okstad, E. Helgason, J. Rilstone, M. Wu, J.F. Kolonay, M.J. Beanan, R.J. Dodson, L.M. Brinkac, M. Gwinn, R.T. DeBoy, R. Madpu, S.C. Daugherty, A.S. Durkin, D.H. Haft, W.C. Nelson, J.D. Peterson, M. Pop, H.M. Khouri, D. Radune, J.L. Benton, Y. Mahamoud, L. Jiang, I.R. Hance, J.F. Weidman, K.J. Berry, R.D. Plaut, A.M. Wolf, K.L. Watkins, W.C. Nierman, A. Hazen, R. Cline, C. Redmond, J.E. Thwaite, O. White, S.L. Salzberg, B. Thomason, A.M. Friedlander, T.M. Koehler, P.C. Hanna, A.B. Kolsto, C.M. Fraser, The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria, Nature 423 (2003) 81–86.
[69] W. Kunze, L. Muller, M. Kilian, M.U. Schuhmann, L. Baumann, W. Handrick, Recurrent posttraumatic meningitis due to nontypable *Haemophilus influenzae*: case report and review of the literature, Infection 36 (2008) 74–77.
[70] G. Urwin, J.A. Krohn, K. Deaver-Robinson, J.D. Wenger, M.M. Farley, Invasive disease due to *Haemophilus influenzae* serotype f: clinical and epidemiologic characteristics in the *H. influenzae* serotype b vaccine era. The *Haemophilus influenzae* Study Group, Clin. Infect. Dis. 22 (1996) 1069–1076.
[71] M. Kilian, A taxonomic study of the genus *Haemophilus*, with the proposal of a new species, J. Gen. Microbiol. 93 (1976) 9–62.
[72] T.F. Murphy, A.L. Brauer, S. Sethi, M. Kilian, X. Cai, A.J. Lesse, *Haemophilus haemolyticus*: a human respiratory tract commensal to be distinguished from *Haemophilus influenzae*, J. Infect. Dis. 195 (2007) 81–89.